



Recipe Recommendation System

Bob Ding

GA DSI 720



Motivations

- Find some thing connected with my teenager daughter.
- Understand more about food and cooking.
- Help people to find new recipes they may like. (quote: Yuval Noah Harari)



Questions to Answer:

1. Given certain ingredients, such as things I have in the fridge, can I find some recipes to tell me what to cook?
2. Given a person's choice of one recipe, could I recommend other similar recipes they might enjoy?
3. Can I categorize the recipes? Given a certain recipe, can I find other recipes that fall into the same categories?

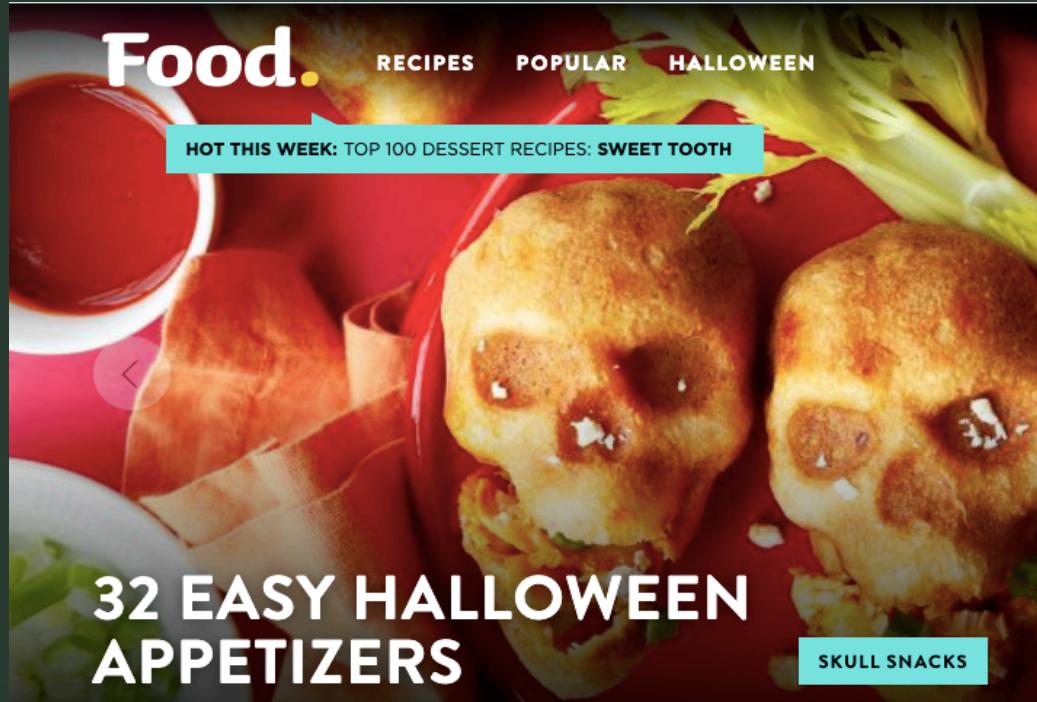
NLP & Recommender in Data Science !

Agenda

- Looking for Recipe Dataset
- Data Cleaning
- Exploratory Data Analysis
- Modeling
- Conclusions and Future Work



Looking for Recipe Dataset



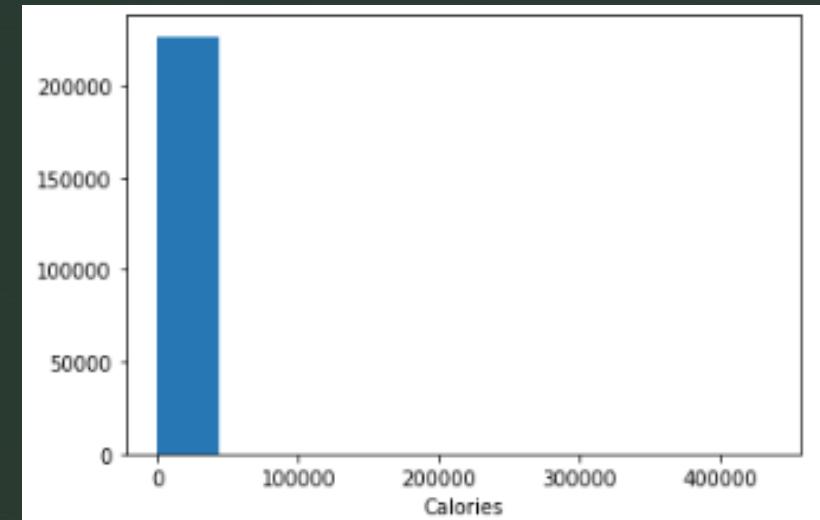
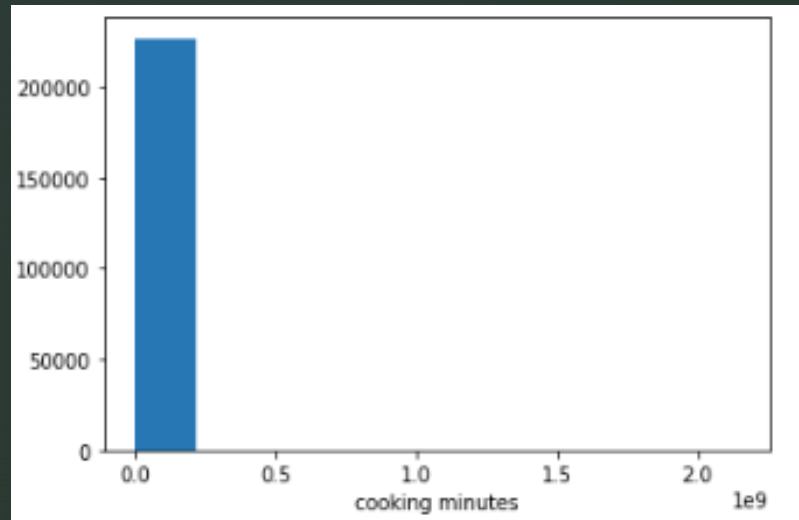
Food.com Dataset stored at Kaggle

230K+ recipes, 1M+ recipe reviews

Data Cleaning

```
recipes.head()
```

	name	id	minutes	contributor_id	submitted	tags	nutrition	n_steps	steps	description	ingredients	n_ingredients
0	arriba baked winter squash mexican style	137739	55	47892	2005-09-16	['60-minutes-or-less', 'time-to-make', 'course...']	[51.5, 0.0, 13.0, 0.0, 2.0, 0.0, 4.0]	11	['make a choice and proceed with recipe', 'dep...']	autumn is my favorite time of year to cook! th...	['winter squash', 'mexican seasoning', 'mixed ...']	7



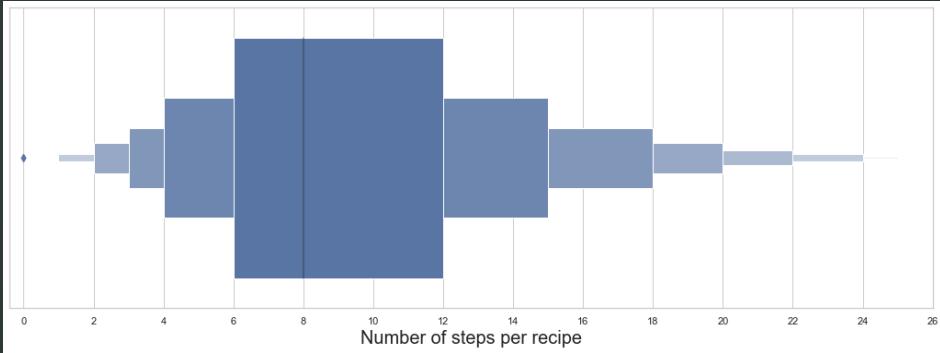
Drop rows of missing values and extreme values, return 191,481 recipes

```
Ratings.head(1)
```

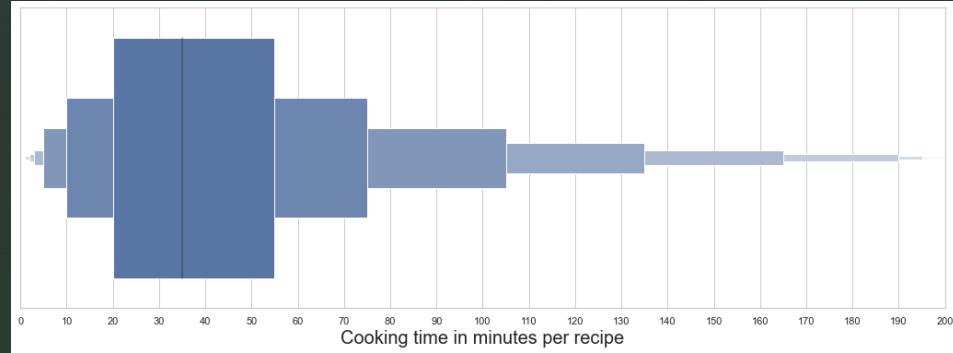
	user_id	recipe_id	date	rating	review
0	38094	40893	2003-02-17	4	Great with a salad. Cooked on top of stove for...

Drop rows of missing values and 0 ratings, return 1,071,351 reviews

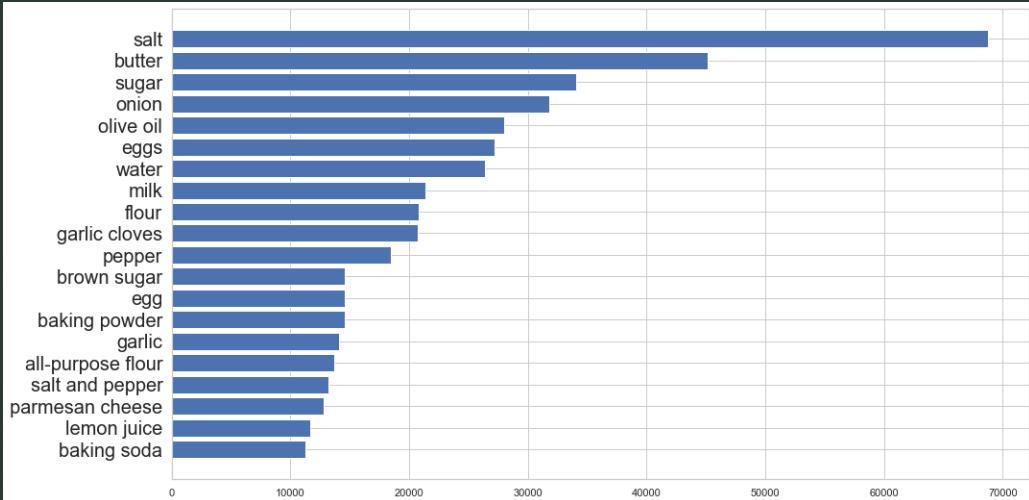
Exploratory Data Analysis



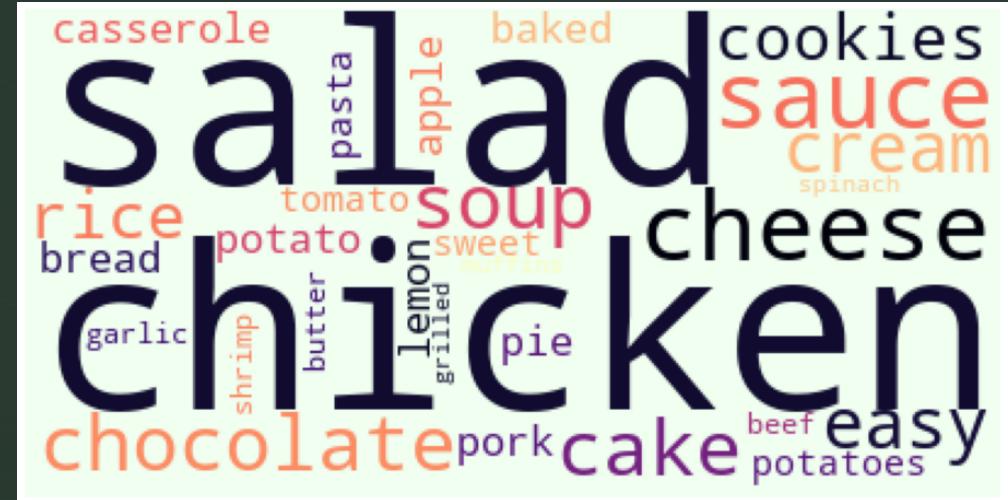
Top ingredients



Top Names

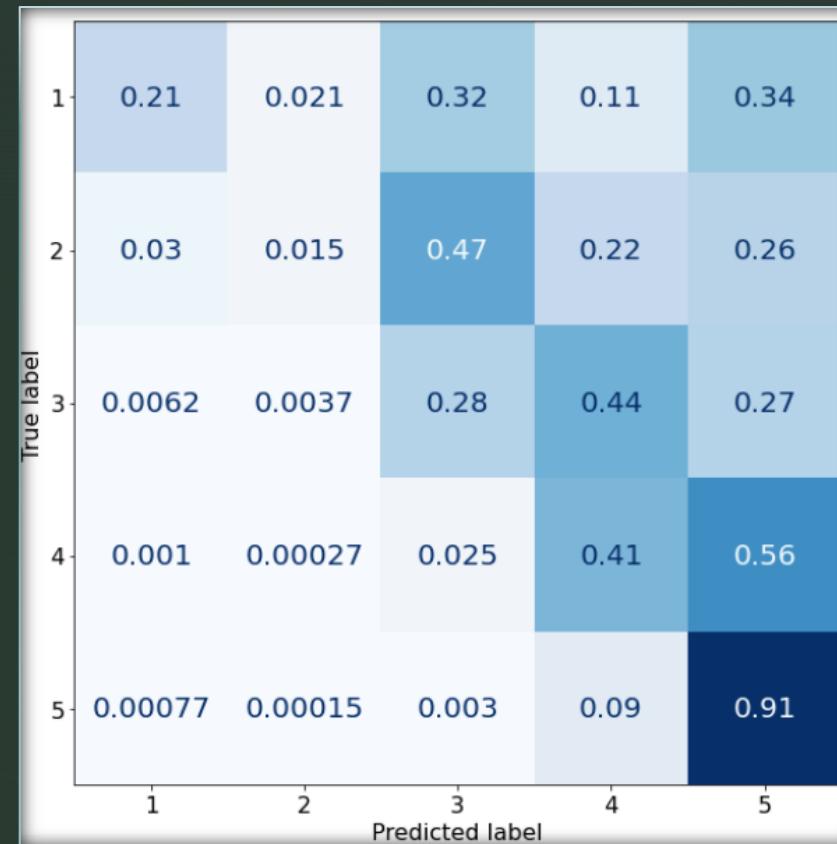


Top ingredients



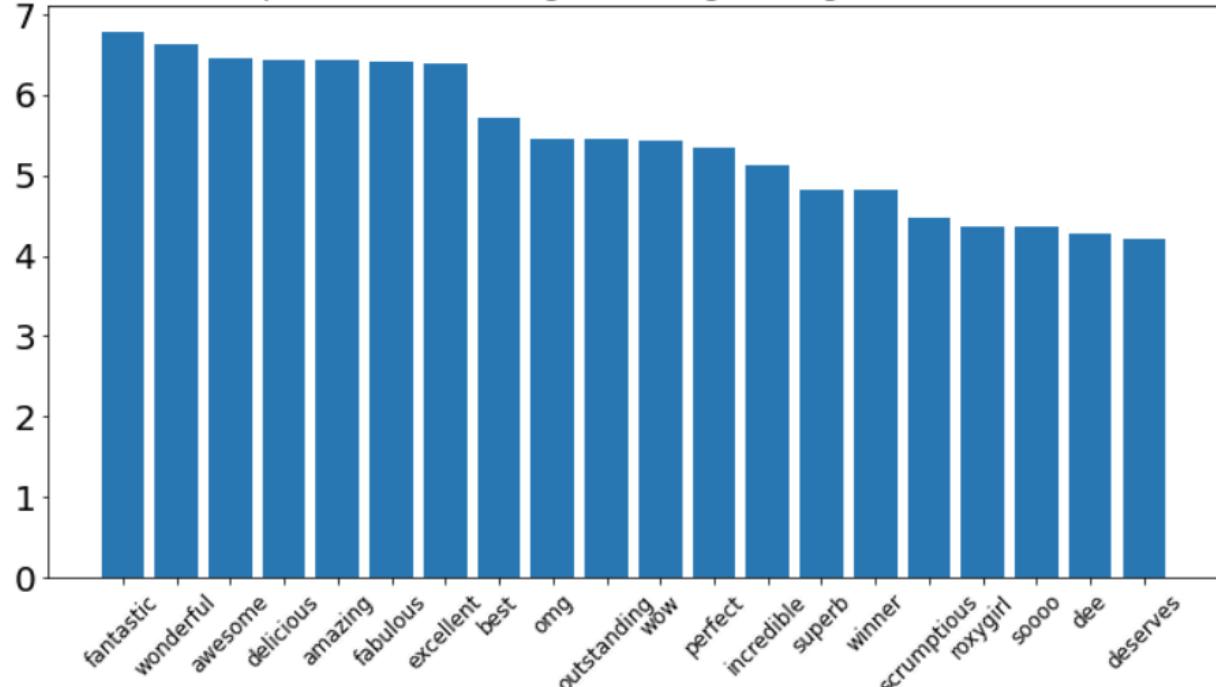
Model 1: Classification for Review Text with NLP

- Features: User review text.
- Vectorizer: CountVectorizer, TfidfVectorizer
- Classifier: MultinomialNB, GradientBoost, LogisticRegression
- Labels: User rating 1 to 5

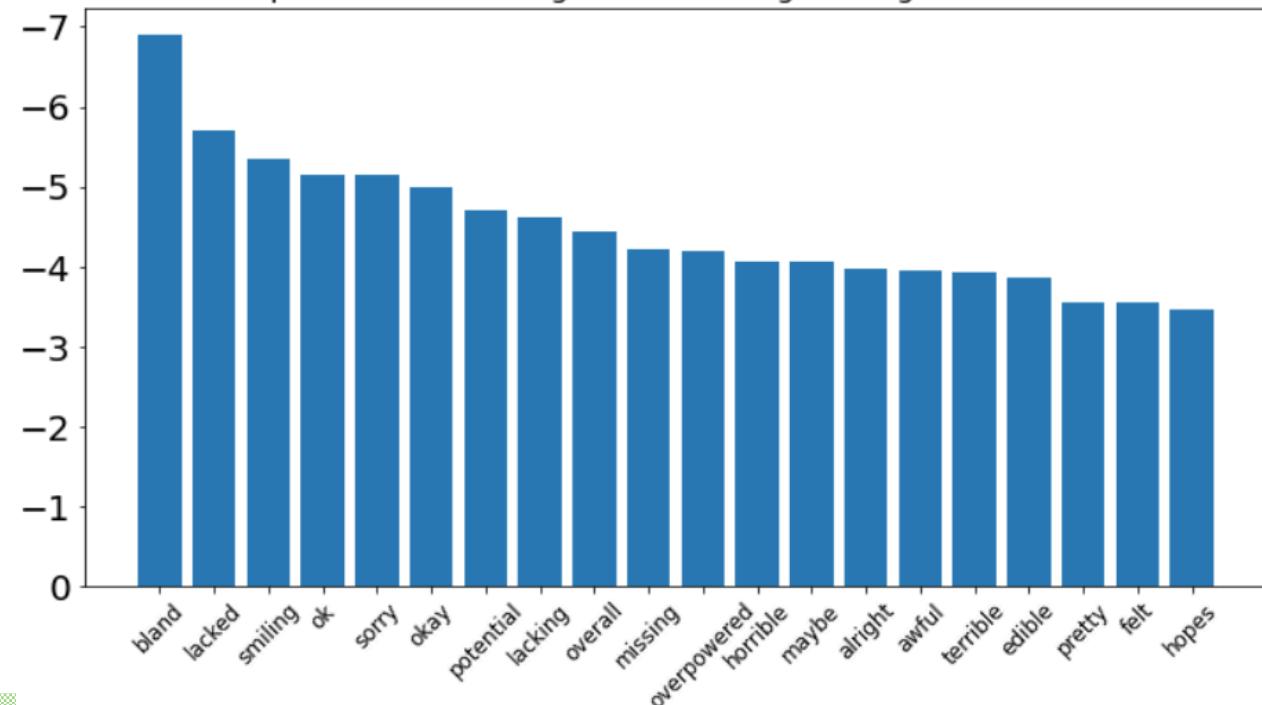


Confusion Matrix of the best estimator from Pipeline and GridSearch

Top 20 Words for rating 5 from Logistic Regression Model



Top 20 words for rating 1 to 4 from Logistic Regression Model



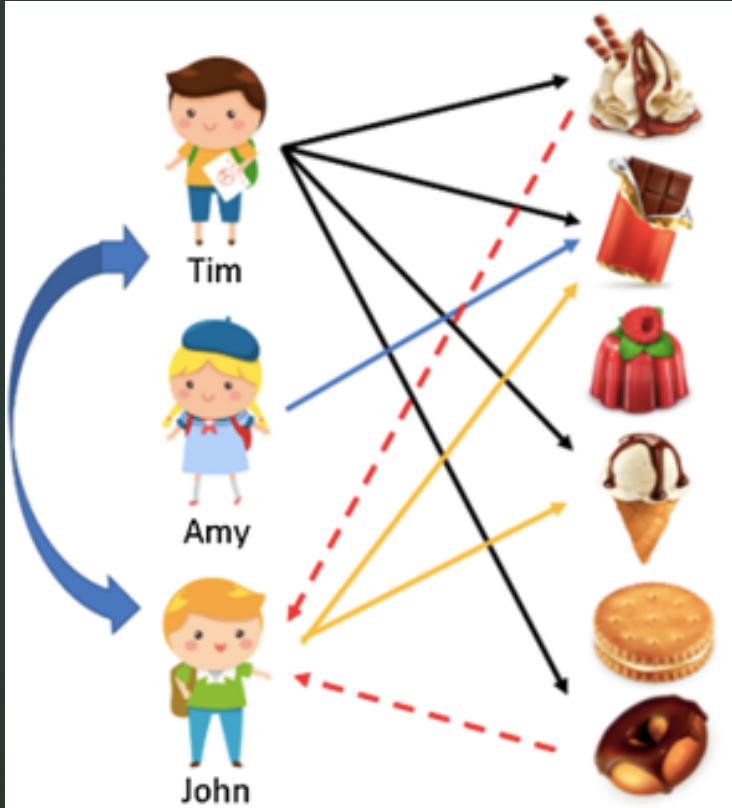
Model 2: Recipe Matching

If I have certain ingredients in my fridge, such as winter squash, Mexican seasoning, mixed spice and honey, the model can give recipes for what to cook with these ingredients.

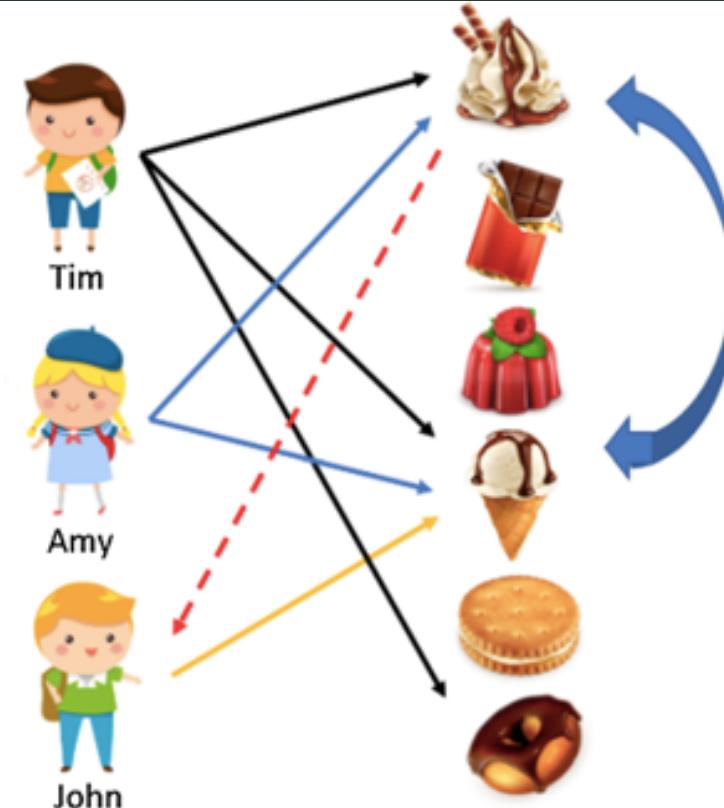
query = 'winter squash, Mexican seasoning, honey, butter, salt, olive oil'

name	minutes	ingredients	steps	match
arriba baked winter squash mexican style	55	[winter squash, mexican seasoning, mixed spice, honey, butter, olive oil, salt]	['make a choice and proceed with recipe', 'depending on size of squash , cut into half or fourth...']	6
skillet pork chop saute with peaches	20	[olive oil, boneless pork loin chops, salt, black pepper, shallots, fresh thyme, peaches, dry wh...	['heat a large skilletover medium-high heat , add oil to pan , and swirl to coat', 'sprinkle cho...	4
anne s garlic knots	150	[warm water, yeast, honey, extra virgin olive oil, salt, flour, parmesan cheese, butter, garlic,...	['proof yeast in water , honey and oil', 'let sit for about 5 minutes or until foamy', 'once pro...	4

Model 3: Collaborative Filtering

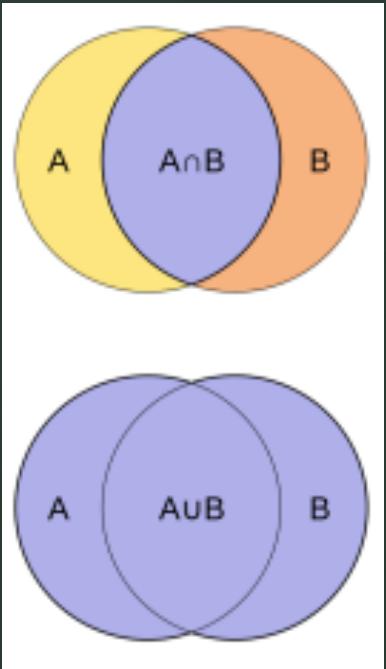


(a) User-based filtering



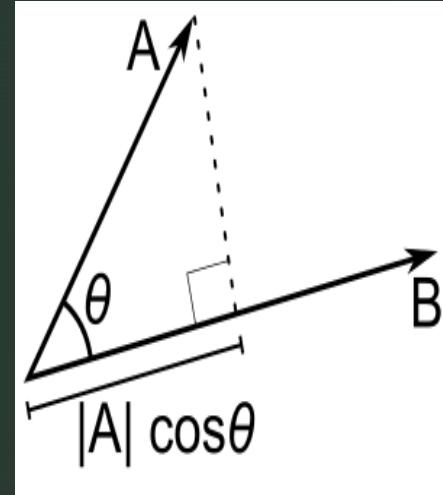
(b) Item-based filtering

Model 3: Collaborative Filtering—math behind



$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$

Jaccard Similarity
Treat rating as set



$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|}$$

Cosine Similarity
Treat rating as vector

Model 3: Collaborative Filtering -- Testing

Item based collaborative filter

query = 'fried chicken'

name	
4th of july fried chicken	1.000000
deep in the heart of texas bbq rub	0.432338
spicy egg salad sandwiches	0.406579
hawaiian ham and swiss sandwich	0.400000

name	
vegan fried chicken vegan chicken nuggets gluten free	1.000000
besan chickpea flour pastry	0.256495
good old fashioned english chip shop style chips	0.247537
garam masala green beans	0.241684
vegan speedy alfredo style sauce no tofu	0.234146
southwestern sugar cookies	0.232119

User based collaborative filter

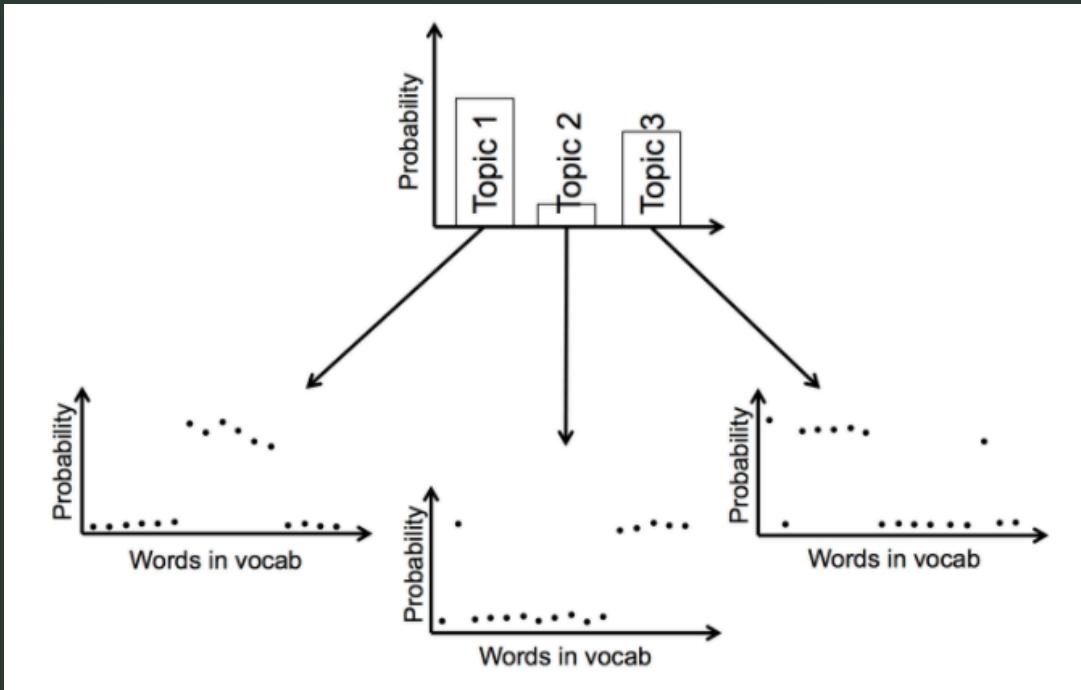
query = '2118338'

user_id	
547623	1.000000
2118338	1.000000
1580557	0.514496
796254	0.478913
88378	0.145371
1271905	0.000000

Model 4: Content Based Filtering



Latent Dirichlet Allocation(LDA)



Unsupervised learning

Model 4: Content Based Filtering--Testing

Number of topics: 30

Corpus: bag of word by ingredient

query = 'winter squash, Mexican seasoning, honey, butter, salt, olive oil'

query_vector = [(13, 0.48639968), (24, 0.13133611), (27, 0.26975682)]

top 5 recipe with index number and similarity

(0, 1.0),
(91318, 0.9725976),
(150490, 0.97194046),
(67211, 0.9719329),
(143749, 0.9719295)]

	name
0	arriba baked winter squash mexican style
91318	german erdbeerbowl strawberry wine punch
150490	outback croutons
67211	crunchy garlic croutons

Conclusions and future work

- Model 2, 3 and 4 recommender works well when input query separately on Jupiter notebook.
- Need more work to combine the simple search, collaborative filtering and content-based filter together.
- Web app (Flask) works and return results. But need more work to improve user interface and shorten the running time.
- Generate new recipe?

Acknowledgements

- Microsoft PowerPoint for slides design
- Elizabeth Ding's blog <https://paintpencilpastries.com/> for food photos
- Kelly S. and Noah C, private communications.